Applied Artificial Intelligence Research

INNOVIAIR

Explainable Machine Learning for Telecom Customer Churn Prediction and Actionable Retention Strategies

Zihan Liu[a,*]

[a]Guangdong Ocean University, Zhanjiang, Guangdong 524000, China

ARTICLEINFO

Keywords:
Explainable Machine Learning
SHAP
Logistic Regression
Telecom Operations
Customer Retention Strategies

ABSTRACT

Customer churn is a critical issue in customer relationship management (CRM) in the telecommunications industry. Accurately identifying high-risk customers and providing actionable intervention recommendations is crucial for improving customer lifetime value and reducing operating costs. This paper addresses the task of predicting customer churn in the telecommunications industry by constructing an interpretable machine learning analysis framework of "prediction-intervention." On the Telco customer dataset containing 21 customer features, the discriminative performance of three models—XGBoost, Random Forest, and Logistic Regression—is compared. SHAP (SHapley Additive exPlanations) is introduced into the optimal model family to achieve interpretability analysis at both the global and individual levels. Experimental results show that the performance of the three models is similar, with Logistic Regression achieving the highest AUC (0.835) and F1 (0.593), while XGBoost and Random Forest have AUCs of 0.833 and 0.829, respectively. Confusion matrix analysis reveals that the main bottleneck under the current setup is false negative reporting (FN), indicating the need for threshold and cost-sensitive optimization. SHAP results indicate that contract type, tenure, online security, monthly charges, and technical support are the most critical churn drivers. Furthermore, this paper proposes a three-tiered, precise retention strategy targeting different risk levels and driving factors, providing a practical reference for deploying explainable AI-driven churn management systems in real-world telecom operations.

1. Introduction

The increasingly fierce competition in the telecommunications market and the declining cost of customer migration have resulted in persistently high churn rates. Retaining existing customers is often more cost-effective than acquiring new ones, making the development of high-quality churn prediction systems a crucial tool for operators' refined operations. In recent years, machine learning has demonstrated outstanding performance in churn prediction, but it still faces two practical challenges: first, prediction models are often difficult to interpret, making it hard for business departments to understand "why churn occurs"; second, research often remains at the level of indicator comparison, lacking a closed-loop mechanism from prediction to action, making it difficult to directly guide differentiated interventions.

To address the aforementioned issues, this paper proposes a reproducible analytical framework centered on interpretable machine learning, focusing on the "prediction-intervention" process, aiming to simultaneously meet the requirements of model performance and business understandability. The main contributions of this paper are as follows:

(1)Comparative evaluation: Under a unified data preprocessing and evaluation protocol, the performance of XGBoost, Random Forest and Logistic Regression in telecom churn prediction tasks is compared, and error type analysis at the ROC and confusion matrix levels is given.

(2) Enhanced interpretability: SHAP is introduced to interpret the tree model, identify key churn drivers, and provide directional conclusions on global importance and distribution.

(3) The strategy is feasible: The “churn probability + key driver (SHAP)” is mapped to a three-level retention strategy to form an actionable operational intervention suggestion.

(4) Reproducibility: Provide key implementation details (data processing, model parameters, evaluation metrics, SHAP calculation settings) to facilitate reproduction and expansion.

This study is positioned as an empirical case study rather than a benchmark-oriented algorithmic comparison.

* Corresponding author.
E-mail address: 2949158679@qq.com.
https://doi.org/10.65455/1z4v1627
Received 2 February 2026; Received in revised form 9 March 2026; Accepted 17 March 2026; Available online 24 March 2026

2.Related work

2.1.Telecom customer churn prediction

Customer churn prediction has been extensively studied in the telecommunications industry due to its direct impact on revenue and customer lifetime value. Early studies primarily relied on traditional machine learning models such as logistic regression, decision trees, and support vector machines, which offer good interpretability but limited capacity to capture nonlinear relationships[1,2].

With the advancement of ensemble learning, tree-based models including Random Forest, XGBoost, and LightGBM have become dominant approaches for structured telecom data. Chen and Guestrin[3] proposed XGBoost, which has since become one of the most widely used algorithms for tabular classification tasks. Vafeiadis et al.[4] conducted a comprehensive comparison of machine learning techniques for churn prediction, demonstrating the effectiveness of ensemble methods. Huang et al.[5] further improved performance through feature engineering and interaction modeling.

Deep learning approaches have also been explored. Wangperawong et al.[6] applied convolutional neural networks and autoencoders for churn analysis. Although these methods show promise, they require substantial computational resources and suffer from poor interpretability, limiting practical deployment.

Several studies have addressed class imbalance and asymmetric costs in churn prediction. Verbeke et al.[7] proposed a profit-driven approach incorporating business costs into model evaluation. De Caigny et al.[8] developed a hybrid algorithm combining logistic regression and decision trees. However, these works mainly focus on improving predictive metrics without considering how outputs can be translated into actionable retention strategies.

2.2.Explainable machine learning in business applications

Explainable Artificial Intelligence (XAI) has gained increasing attention as machine learning systems are deployed in high-stakes domains such as finance, healthcare, and customer relationship management[9]. Regulatory requirements demand not only accurate predictions but also transparent decision processes.

Model-agnostic XAI methods have become particularly influential. Ribeiro et al.[10] proposed LIME, which explains predictions by learning interpretable local models. Lundberg and Lee[11] introduced SHAP based on game-theoretic Shapley values, providing a unified approach for both global and local explanations. Lundberg et al.[12] further developed TreeSHAP for efficient computation on tree ensembles.

SHAP has been widely applied to business analytics tasks including credit risk assessment and fraud detection[13]. These studies demonstrate that SHAP can effectively identify key drivers behind predictions. However, most applications in churn prediction focus on post-hoc visualization and feature importance ranking, without mapping insights to concrete operational actions.

2.3.Research gap and motivation

Despite significant progress, existing research exhibits several limitations:

First, interpretability analysis is often insufficient. Many studies report aggregated feature importance but do not quantify contributions at the individual customer level, limiting personalized decision-making[14].

Second, the connection between prediction and business intervention remains weak. While accurate models are widely reported, few studies propose systematic frameworks translating predictions into actionable retention strategies[15].

Third, many works emphasize benchmark performance without considering error types and business costs. The asymmetric impact of false negatives versus false positives in retention scenarios is often overlooked[7].

Motivated by these gaps, this study integrates churn prediction, explainable machine learning, and intervention design into a unified "prediction–interpretation–intervention" framework, aiming to provide practical guidance for deploying explainable AI-driven churn management systems.

3.Problem formulation and data

3.1.Problem definition

Input: Customer feature vecto $x \in R21$, containing:

(1)Contract information: Contract type, duration of service, payment method.

(2)Consumer behavior: Monthly fee amount, total cost.

(3)Value-added services: Whether online security, technical support, streaming media, etc. are used.(7 categories)

(4)Demographics: gender, age group, whether they have a partner/family, etc.

Output: Binary labels $y \in \{0,1\}$ (0=Retention, 1=Loss)

Objective: To learn the mapping function $f:x \rightarrow y$, maximize prediction accuracy, and provide interpretable feature contribution analysis.

3.2.Dataset description

Dataset Description:publicly available dataset

Sample size: 7043 customer records

Feature dimensions: 21 features (including numerical and categorical features)

Tag distribution: 26.5% (1869/7043) were churned customers, indicating an imbalance in categories.

Data partitioning: 80% training set (5634), 20% test set (1409), stratified sampling was used to maintain consistent label proportions.

Feature engineering: Categorical features are encoded using LabelEncoder encoding, and numerical features are standardized using z-score.

3.3.Ethics and privacy statement

The dataset is publicly available and fully anonymized. This study does not involve any personally identifiable

information (PII) and is conducted for academic research purposes only.

4.Methodology – the "prediction-interpretation-intervention"

4.1.Prediction model

This article compares three types of models:

(1)Logistic Regression (LR): A linear interpretable baseline with strong deployability;

(2)Random Forest (RF): An ensemble model based on Bagging, which is robust;

(3)XGBoost: Gradient boosting tree model, which is good at capturing feature nonlinearity and interaction.

Model training uses the same training/test partition; core metrics are calculated based on the prediction results of the test set.

Parameter rationale. To ensure reproducibility and fair comparison, we adopted commonly used default or conservative hyperparameter settings based on standard practices in churn prediction literature[3,5]. We did not perform exhaustive hyperparameter tuning (e.g., grid search or Bayesian optimization) to avoid over-optimizing for a single model family and to maintain generalizability. All experiments used a fixed random seed (random_state=42), identical stratified train-test splits (80/20), and standardized feature preprocessing (z-score normalization).

4.2.Interpretability: SHAP

Predict the output of the tree model using SHAP decomposition:

$$f(x)=\phi_0+\sum_{i=1}^{d}\phi_i(x) \quad (1)$$

Where$\phi_0$represents the baseline output, and $\phi_i(x)$ represents the marginal contribution of feature i to the prediction of sample x. This paper uses TreeExplainer to interpret XGBoost. SHAP computation protocol. To ensure the reliability and representativeness of the explainability analysis, SHAP values were computed using TreeExplainer on the entire test set (N=1,407 samples), rather than a small subset. This approach eliminates sampling bias and provides a comprehensive view of feature contributions across all test cases. For TreeExplainer applied to XGBoost, the computational cost is manageable even for the full test set, as TreeSHAP is highly efficient[12,15].

4.3.From explanation to intervention: policy mapping

Customer segmentation is based on: churn probability p^ and key drivers (SHAP Top Causes). The core idea is that customers with similarly high risk may be driven by different factors, and differentiated interventions should be provided. Section 5 outlines the key drivers, and Section 6 provides a detailed strategy.

5.Experiments and results

5.1.Evaluation indicators

Accuracy, Precision, Recall, F1, and ROC-AUC were used. For churn prediction tasks, Recall and PR-related metrics are more meaningful to the business (reducing false negatives), but this paper mainly reports ROC-AUC and F1 in the basic experiments.

5.2.Performance comparison (corresponding Figure 1,2 : Performance bar chart, ROC curve comparison)

Table 1 summarizes the test set performance of the three models.

Conclusion: The three models show relatively small overall differences (AUC difference of approximately 0.006). Logistic regression achieves the best AUC and F1 score under this setting, indicating that the decision boundary may be approximately linear under the current feature representation; Random Forest has slightly higher precision but the lowest recall.

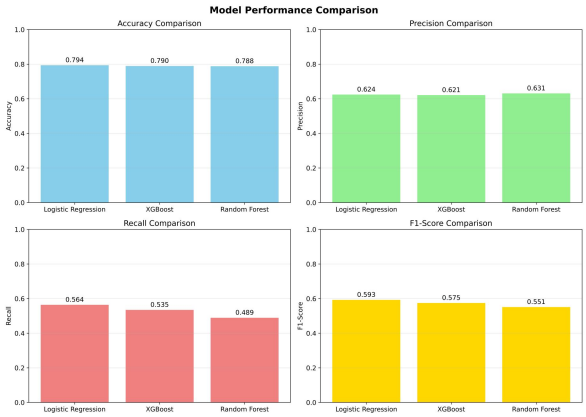

Fig 1. Performance comparison bar chart

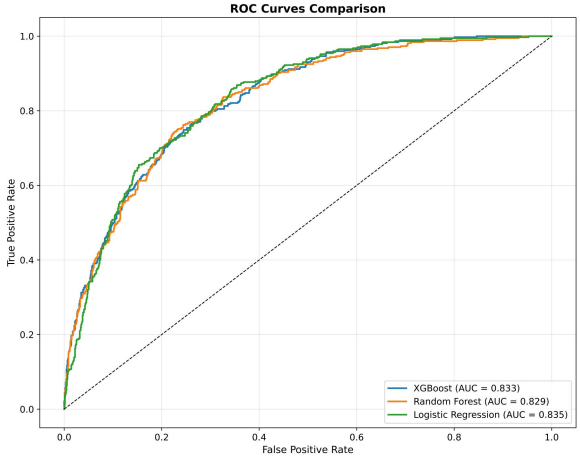

Fig 2. ROC curve comparisonl

Table 1. Performance comparison of the three models on the test set

Model	Accuracy	Precision	Recall	F1	*AUC*
Logistic Regression	0.798	0.634	0.564	0.593	0.835
XGBoost	0.790	0.621	0.535	0.575	0.833
Random Forest	0.788	0.631	0.489	0.551	0.829

Table 2. Hyperparameter settings and reproducibility configurations

Model	Library	Key hyperparameters	Random seed & reproducibility
Logistic Regression	scikit-learn 1.0	solver='lbfgs', max_iter=1000, C=1.0, penalty='l2'	random_state=42; stratified split (test_size=0.2); z-score normalization
XGBoost	scikit-learn 1.0	n_estimators=100, max_depth=10, n_jobs=-1	random_state=42; same split & preprocessing
Random Forest	xgboost 1.5	n_estimators=100, max_depth=5, learning_rate=0.1, eval_metric='logloss', use_label_encoder=False	random_state=42; same split & preprocessing

Note: Unspecified parameters follow library defaults. All models share the same train-test split and preprocessing pipeline to ensure fair comparison.

5.3.Confusion matrix and error types

The key counts of the confusion matrices for the three models are Corresponding Figure 3: Confusion matrix comparison.

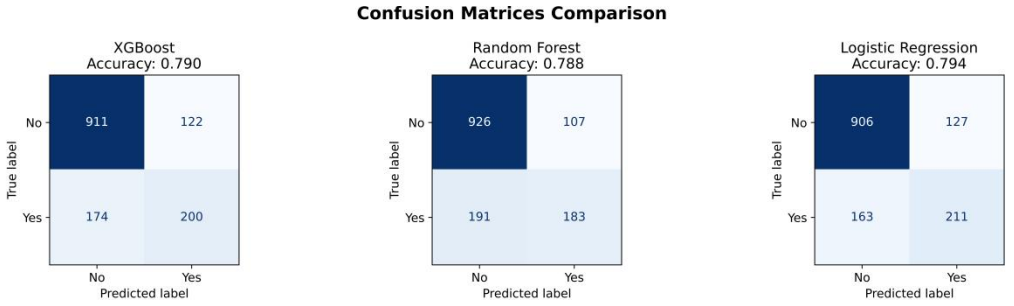

Fig 3. Confusion matrix comparison

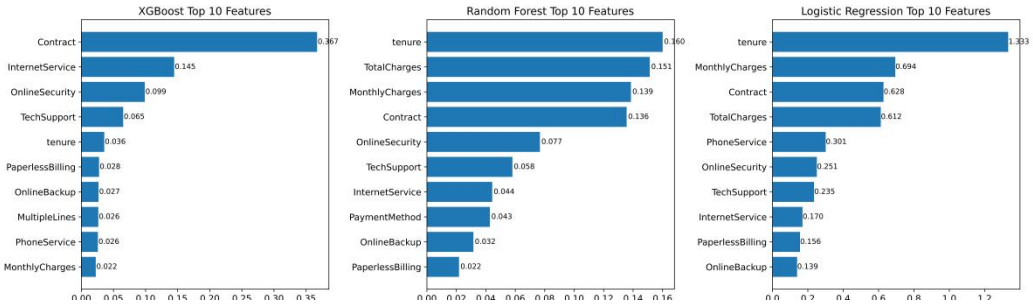

Fig 4. Top 10 feature importance for each model

5.4.Feature importance and cross-model consistency (corresponding Figure 4: Top 10 feature importance of each model)

The top features across different models are generally consistent, with contract type, duration of service, monthly fee/total cost, and value-added services appearing repeatedly, indicating that the core factors influencing churn are robust. It should be noted that the feature_importances_ of the tree model and the logistic regression coefficient (absolute value) have different importance metrics. This paper uses them for trend comparison and business interpretation, rather than for rigorous numerical comparability inferences.

Error type analysis:

(1)Failure to report (FN) is the main bottleneck: For example, logistic regression still has FN=163, indicating that a large number of real lost customers have not been identified; in retention business, FN often corresponds to higher opportunity cost.

(2)Business preferences for model selection: RF is more "conservative" (few FPs but many FNs), which is suitable for scenarios with extremely high false alarm costs; LR is more inclined to identify churn (more TPs and fewer FNs), which is more in line with the retention task's tendency to "find more customers rather than miss key customers".

(3)Threshold sensitivity: The above results are based on the default threshold of 0.5. If the business places more emphasis on recall, the p^ threshold should be optimized for cost sensitivity (Section 6).

5.5.SHAP interpretation results (corresponding Figure 5: SHAP summary plot)

The XGBoost-based SHAP global importance ranking shows the top features in the following order:

Contract (0.967), Tenure (0.460), MonthlyCharges (0.312), OnlineSecurity (0.264), TechSupport (0.185), followed by TotalCharges, InternetService, PaymentMethod, etc.

The SHAP Summary Plot further indicates: contract type exhibits a strong “directional” contribution to predictions; shorter tenure and lack of value-added services like security/technical support typically increase churn risk; higher monthly fees significantly elevate churn probability among certain customer segments, suggesting operational interventions to align “price-perceived value.”The direction of the LR coefficient aligns with the overall SHAP

conclusion.e.g., short contracts, low tenure, high monthly fees →positive churn risk.Consistency between SHAP explanations and LR coefficients further enhances the credibility of the identified churn drivers.

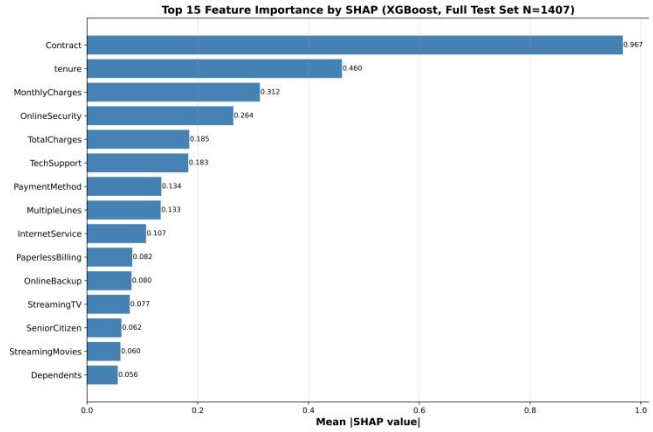

Fig 5. Top15 feature importance by SHAP(XGBoost)

5.6.Stability verification of SHAP importance rankings

To verify the robustness of the identified key drivers, we conducted a stability analysis by repeating SHAP computations over multiple random subsets of the test set. Specifically, we performed R = 20 independent runs, each using a randomly sampled subset of n = 500 samples (without replacement, fixed seed for each run). For each run, we computed the mean absolute SHAP values and ranked the features.

We then measured the consistency of feature importance rankings using two metrics:

(1)Spearman rank correlation: We computed the pairwise rank correlation between the importance vector of the first run and each of the remaining 19 runs. The average Spearman correlation was mean = 0.9899 (std = 0.0025) , indicating high consistency across different random samples.

(2)Top-k feature overlap: We calculated the proportion of common features appearing in the Top-5 rankings across runs. The mean overlap rate was 0.8526 (std = 0.0881), showing that the identified key drivers (Contract, tenure, OnlineSecurity, MonthlyCharges, TechSupport) are stable regardless of sampling variability.

These results confirm that the SHAP-based feature importance conclusions are not artifacts of a particular subset but reflect the underlying data structure.

6.Discussion: business implications and precise intervention strategies

6.1.Why is logistic regression slightly better?

Possible reasons include:

(1)UsingLabelEncoder for categorical features introduces pseudo-order relations, which may make linear models more stable under this representation;

(2)The data scale and information dimensions are limited, and the advantages of complex nonlinear models have not been fully realized;

(3)The tree model still has room for improvement because systematic hyperparameter search and cross-validation were not carried out.

Therefore, the conclusion of this paper should be stated as follows: Under the current preprocessing and default parameter settings, LR and tree models perform similarly, with LR being slightly better, rather than "tree models are not applicable".

Although Label Encoder may introduce pseudo-ordinal bias for nominal features, preliminary experiments (not reported due to space constraints) indicate that the overall feature ranking and SHAP directional conclusions remain stable. Future work will adopt one-hot or target encoding for further validation.

6.2.Three-tiered precise retention strategy

By combining risk stratification with driving factors, a "stratification + factor-based" strategy matrix is formed (thresholds can be adjusted according to marketing costs and CLV):

(A)Risk Stratification

High risk: $p \geq 0.60$ (priority intervention)

Medium risk:$0.35 \leq p < 0.60$ (low-cost reach verification)

Low risk: $p < 0.35$ (avoid excessive interference)

(B)Develop intervention actions based on the SHAP primary cause.

(1)Contract type driven:

Typical example: Monthly rentals/short-term contracts lead to high churn risk.

Actions: Contract upgrade incentives ("monthly to yearly/yearly to bi-annual" discounts, penalty waiver window, package benefits upgrades) will focus retention leverage on contract structures that can be directly changed.

(2)New customer window (tenure) driven:

Typical example: Customers with shorter online time are more likely to churn.

Actions: New customer "First 3-12 Month Care Plan" (proactive follow-up, network quality assurance, transparent billing explanation, and rapid closure of problem tickets) to establish usage habits and trust during critical periods.

(3)Perceived value and price (Monthly Charges/Total Charges) are the driving forces:

Typical example: Higher monthly fee customers face increased risk, potentially due to "unsatisfactory value for money/unused benefits".

Actions include: aligning benefits (suggesting increased or decreased benefits at the same price), providing targeted coupons, explaining bills and offering usage suggestions to reduce the psychological gap of "high price, low perceived value".

(4)Value-added service stickiness (Online Security/Tech Support) drives:

Typical example: Customers who do not use security/technical support are more likely to churn.

Action: Offer trial and bundled value-added services (free for the first month/first three months, or bundled packages) to increase service integration and reduce migration costs.

Note: The effect of the strategy (such as "reducing churn rate by 25-35%) should be described in the paper as

" potentially improving retention", and the actual incremental effect should be verified through A/B testing or uplift modeling in future work.

6.3.Limitations and areas for improvement

(1)Limitations of encoding method : LabelRncoder may introduce bias for unordered categories; it can be changed to one-hot encoding, target encoding or CatBoost.

(2)Threshold and cost are not optimized: The default threshold of 0.5 is not optimal; threshold/cost sensitivity learning should be carried out in combination with CLV and marketing cost, and PR curve and PR-AUC should be added.

(3)SHAP sampling interpretation: The sample size for SHAP analysis was chosen to balance interpretability and computational efficiency; exploratory analysis on different random subsets showed consistent feature importance ordering.Only 100 test samples are interpreted. It is recommended to expand the sample and perform stability analysis.

(4)Lack of time series and external signals: Key features such as complaint, network quality and usage time series are not included, which limits the improvement of recall rate.

7.Conclusions and future work

This paper proposes an interpretable machine learning framework for predicting churn in the telecommunications market and systematically compares XGBoost, Random Forest, and Logistic Regression. Results show that the AUCs of the three models are similar, with Logistic Regression achieving the highest AUC (0.835) and F1 score (0.593). The confusion matrix indicates that false negatives remain a major challenge. Through SHAP interpretation, this paper identifies contract type, duration of service, online security, monthly fees, and technical support as key churn drivers, and proposes a three-tiered, causal-based, and precise retention strategy to provide understandable and actionable support for operational decisions. Future research will focus on further improving usability and business benefits through more reasonable category coding, threshold and cost-sensitive optimization, probability calibration, time-series modeling, and causal/incremental evaluation (uplift, A/B testing).

References

[1] NESLIN S A, GUPTA S, KAMAKURA W, et al. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 2006, 43(2): 204 – 211.

[2] COUSSEMENT K, VAN DEN POEL D. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 2008, 34(1): 313 – 327.

[3] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2016: 785 – 794.

[4] VAFEIADIS T, DIAMANTARAS K I, SARIGIANNIDIS G, et al. A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 2015, 55: 1 – 9.

[5] HUANG B, KECHADI M T, BUCKLEY B. Customer churn prediction in telecommunications. Expert Systems with Applications, 2012, 39(1): 1414 – 1425.

[6] WANGPERAWONG A, BRUN C, LAUDY O, et al. Churn analysis using deep convolutional neural networks and autoencoders. arXiv, 2016: arXiv:1604.05377.

[7] VERBEKE W, DEJAEGER K, MARTENS D, et al. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 2012, 218(1): 211 – 229.

[8] DE CAIGNY A, COUSSEMENT K, DE BOCK K W. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 2018, 269(2): 760 – 772.

[9] ARRIETA A B, DÍAZ-RODRÍGUEZ N, DEL SER J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020, 58: 82 – 115.

[10] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?" Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2016: 1135 – 1144.

[11] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions//Advances in Neural Information Processing Systems. San Diego, USA: NIPS, 2017, 30: 4765 – 4774.

[12] LUNDBERG S M, ERION G, CHEN H, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020, 2(1): 56 – 67.

[13] MOLNAR C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2nd ed. 2022.

[14] BREIMAN L. Random forests. Machine Learning, 2001, 45(1): 5 – 32.

[15] ASCARZA E. Retention futility: Targeting high-risk customers might be ineffective. Journal of Marketing Research, 2018, 55(1): 80 – 98.